Int J Artif Intell Educ (2017) 27:393–418
DOI 10.1007/s40593-016-0129-5

ARTICLE

Role of Working Memory and Strategy-Use in Feedback Effects on children's Progression in Analogy Solving:an Explanatory Item Response Theory Account

Claire E. Stevenson [1,2]

Published online: 20 December 2016

Abstract This study contrasted the effects of tutoring, multiple try and no feedback on children's progression in analogy solving and examined individual differences herein. Feedback that includes additional hints or explanations leads to the greatest learning gains in adults. However, children process feedback differently from adults and effective feedback likely differs between learners with different characteristics or at different stages in the learning process. In this paper multilevel explanatory item response theory models were used to examine individual differences in feedback effects in children's performance on a computerized pretest-training-posttest assessment of analogical reasoning. The role of working memory and ability level, based on initial strategy-use, were examined in a sample of 999 5–10 year-old children who received either tutoring feedback, multiple tries or no feedback during the training sessions. The results indicate that tutoring feedback leads to the greatest performance gains; however, this was moderated by working memory and ability level. Children who initially used less advanced strategies benefited more from each type of feedback than children who used advanced strategies at pretest. Higher working memory scores were linked to greater benefit from tutoring feedback or no feedback, whereas learning gains in the multiple try condition were not related to working memory. The findings of this study contribute to the growing literature on how to personalize feedback to the learner's instructional-needs.

Keywords Feedback effects · Elaborate feedback · Knowledge of response · Knowledge of correct response · Figural analogies · Measuring change · Explanatory IRT

✉ Claire E. Stevenson
c.e.stevenson@uva.nl

[1] Leiden University, Leiden, Netherlands

[2] Psychological Methods Department, University of Amsterdam, Postbus 15906, 1001NKAmsterdam, The Netherlands

Introduction

Nearly a century of feedback research has demonstrated large variation in feedback effects on learning (Kluger and DeNisi 1996); this implies that individual learners may react differently to different types of feedback in different contexts. Recent research shows that digital feedback effects depend on learner characteristics (e.g., Narciss 2004); this is even the case during adaptive hint procedures in intelligent tutoring systems (Goldin and Carlson 2013; Narciss et al. 2014a, 2014b). It is plausible that feedback, i.e. "actions taken by external agents to provide information regarding aspects of one's task performance" (Kluger and DeNisi 1996, p. 255), tailored to learner characteristics may lead to greater learning gains than feedback that is not personalized (e.g., Narciss et al. 2014a, 2014b). However, in order to provide personalized feedback large-scale studies need to be conducted in target populations to provide evidence-based guidelines concerning which type of feedback is optimal for different types of learners.

In digital learning environments different types of feedback can be used. Shute (2008) distinguished a range of feedback-types from simple forms such as response verification to elaborate feedback where errors may be flagged, explanations provided and/or strategic prompts are given on how to proceed with the problem. Elaborate feedback, an umbrella term for any form of simple feedback that provides additional hints or explanations, is generally considered more effective than outcome feedback because it provides specific information beyond correctness and a clear direction of how to proceed and likely increases motivation (Narciss, 2008; 2012; Shute 2008). A meta-analysis of the learning effects of item-based feedback in computer-based environments reported higher effect sizes for elaborate feedback than simple feedback, especially in higher-level learning outcomes where transfer of previous learning to new situations or tasks is required (Van der Kleij et al. 2015). However, the effectiveness of simple versus elaborate feedback may also depend on learner characteristics. Learner level and ability especially appear to play an important role in the effect of different forms of feedback on learning (Hattie and Timperley 2007); lower ability learners may not have the (meta-) cognitive skills to rethink their solutions after receiving outcome feedback whereas higher ability learners can draw on previous knowledge to come up with an alternative (Mason and Bruning 2001; Narciss and Huth 2004). Working memory efficiency also influences one's ability to process feedback as benefitting from feedback is dependent upon successful goal-directed search and retrieval of relevant information from memory (Bangert-Drowns et al. 1991). Working memory may be especially influential in children's feedback processing as working memory capacity is still developing (Alloway et al. 2004). Only a handful of studies have investigated digital feedback effects specifically in primary school children (e.g., Narciss and Huth 2006; Kramarski and Zeichner 2001); as such children form an underrepresented group in the computerized feedback literature (Van der Kleij et al. 2015). Furthermore, children appear to process feedback differently from adults (Eppinger et al. 2009), perhaps in part due to developmental changes in brain regions that govern working memory efficiency (Van Duijvenvoorde et al. 2008). Therefore, it is questionable whether findings in the feedback literature generalize to this age group. Given increased usage of computer-based adaptive materials in primary education (National Academy of Education 2013) it is important to investigate which factors play a role in children's

ability to learn from feedback in digital learning environments. Here we examine the role of individual differences in working memory and prior ability on the effect of feedback on children's analogy solving progression in a digital learning environment.

Analogical reasoning is the cognitive process of recognizing relationships and then transferring information from a known source to a new but similar context that develops with great variability in childhood (Gentner 1983; Siegler and Svetina 2002). For example, a classical visual analogy is "○ is to ● as □ is to ?". Analogical reasoning forms the core of human cognition and intelligence (Sternberg and Gardner 1983). Furthermore, analogies are pervasive in education as teaching tools (e.g., Dagher 1995; Richland et al. 2004) and deemed essential to school learning (Goswami 1991).

Given the essential role analogical reasoning plays in learning (e.g., Goswami 1991), it is important to investigate which type of feedback best promotes analogy solving for individual children in specific learning contexts. Previous research reveals that outcome feedback is more effective than practice alone (e.g., Cheshire et al. 2005). Also, elaborate feedback comprising increasingly specific strategic hints – referred to as tutoring feedback (c.f. Narciss 2013), appears to lead to greater performance gains than simple feedback (Stevenson et al. 2013b). On the whole, providing young children with feedback improves analogical reasoning, where duplication errors, in which one of the analogy terms is copied, decrease and partial and correct analogical solutions increase (e.g., Cheshire et al. 2005; Siegler and Svetina 2002).

A child's ability to process and learn from feedback as well as solve analogies likely depends on his/her working memory efficiency (Stevenson et al. 2013a). Numerous studies have demonstrated a relationship between working memory and analogical reasoning ability in children (e.g., Alloway et al. 2004; Stevenson et al. 2013a; Thibaut et al. 2010). Working memory also appears to moderate the development of analogical reasoning (Richland and Burchinal 2013; Thibaut and French 2016). However, the role of working memory on children's change in analogical reasoning after receiving feedback is unclear and studies up until now have produced conflicting results (e.g., Stevenson et al. 2013a; Stevenson et al. 2013b). In this study the focus is on simple feedback versus elaborate feedback as these two types of feedback are often contrasted in the literature and elaborate feedback provides clear advantages for adults. However, now the specific research question concerns the moderating role of working memory on their effects.

The role of working memory in feedback processing is difficult to predict. On the one hand, the elaborate feedback applied in this study uses a tutoring feedback strategy (e.g., Narciss and Huth 2004; Narciss 2013) that builds up from general to specific hints that guide the learner to the correct solution with knowledge of response (KR) at each step along the way. This carefully designed series of graduated prompts aimed to lessen the cognitive load by providing step-by-step information on the goal of the task, which features were relevant to achieving this goal and if needed scaffolds to help the child solve the task (Campione and Brown 1987; Resing and Elliott 2011). Previous studies support this as diverse populations including children with learning and developmental disabilities – often accompanied by working memory deficits – appear to benefit from graduated prompting techniques (e.g., Resing et al. 2012). On the other hand, elaborate feedback by definition provides more information than simple feedback. This means that more information (although step-wise, specific and directional) in addition to KR

must be processed. Given that analogy solving in itself taxes working memory the processing demand may be too high to provide additional benefit – above that of simple feedback – for children with limited working memory capacity. Support for this line of reasoning can be found in the results of Fyfe and Rittle-Johnson (2016) who investigated the moderating role of working memory on feedback in 7–9 year-olds learning and transfer of mathematics. Their findings indicate that children with greater working memory benefit similarly from feedback on the strategy the child used or simple feedback (knowledge of correct response, KCR), whereas children with lower working memory scores derived the most benefit from KCR. This was tentatively discussed in terms of cognitive load theory (Sweller 1994), where strategy feedback was considered more cognitively demanding.

A second source of individual differences this study examined is the role of learner ability level. Initial analogical reasoning skill appears to be a good predictor of the effect of different forms of feedback on learning. For example, a previous study on children's change in analogical reasoning found that graduated prompts led to greater performance gains on the whole in comparison to multiple try feedback; in addition, this form of interactive tutoring feedback was most effective for children who performed poorly on the pretest (Stevenson et al. 2013b). A possible implication is that providing elaborate rather than simple feedback is not necessarily more beneficial for advanced learners (e.g., Hanna 1976). However, given the intertwined roles of age and working memory in learning to solve analogies it is important that an examination of the role of prior ability on feedback effects controls for differences in age and working memory scores.

In analogical reasoning, strategy-use is related to ability level. Analogy solution strategies can generally be categorized into two groups: analogical versus non-analogical. Analogical reasoning strategies are those in which the child integrates information from each of the elements and relations in the presented problem and imply that the child understands how to solve the task, although errors may be made (Stevenson et al. 2013a). Non-analogical reasoning strategies are associative solutions, such as duplicating one of the analogy terms, or idiosyncratic solutions (Siegler and Svetina 2002; Stevenson et al. 2016). Based on earlier feedback literature of the role of ability level (e.g., Shute 2008; Hattie and Timperley 2007), learners who already apply analogical reasoning strategies, i.e. who have automated the correct strategy to solve analogies, would be expected to benefit regardless of feedback-type. However, children who use less effective strategies (non-analogical strategies) may require elaborate feedback to progress (e.g., Clariana 1990; Hanna 1976). In the present study this was examined by comparing feedback effects on the learning of children who initially used analogical versus non-analogical reasoning strategies while controlling for age and working memory differences.

Current Study

Research has revealed that different types of feedback can improve children's learning of analogical reasoning in general and certain learner characteristics (e.g., age, prior ability, learning phase, motivation, working memory) may influence the degree to which feedback interventions are effective. This research questions this study addressed were:

1. Which type of feedback (elaborate using interactive tutoring feedback, simple multiple try feedback or no feedback) leads to the greatest learning gains in children?
2. Are feedback effects moderated by working memory capacity – i.e. do children with different levels of working memory benefit differently from different types offeedback?
3. Are feedback effects moderated by prior ability determined by initial strategy-use?

To this end a pretest-training-posttest design was employed and children were trained in analogy solving using either interactive tutoring feedback (a form of elaborate feedback), multiple try feedback (a form of simple feedback, Van der Kleij et al. 2015) or practice without feedback and assessed their ability and working memory prior to training. Based on the findings discussed above we expected that interactive tutoring feedback would lead to greater progression in analogical reasoning than multiple try feedback and that both forms of feedback would be more effective than no feedback (hypothesis 1). We expected this would be moderated by individual differences in working memory capacity (hypothesis 2), but did not have prior expectations about the direction of this moderation effect as both the theory and experiment results are complex and contradictory. Children with initially less effective strategies were expected to benefit more from feedback, especially interactive tutoring feedback, in comparison to children who were already capable of applying analogical reasoning strategies after controlling for the effects of age and working memory (hypothesis 3).

Methods

Participants

Participants were 999 children from five age-groups (kindergarten, first through fourth grade) recruited from 26 public elementary schools of similar middle class SES in the south-west of the Netherlands. The sample consisted of 374 boys and 625 girls, with a mean age of 7 years, 3 months (range 4.9–11.3 years). The schools were selected based upon their willingness to participate and written informed consent for children's participation was obtained from the parents.

Design & Procedure

The data utilized in this study is a combination from six separate studies utilizing a pretest-intervention-posttest control-group design.[1] In each study the children were assigned to the interactive tutoring feedback (graduated prompts), multiple-try feedback or a control condition without feedback by blocking based on their scores on a cognitive ability reasoning subtest (visual exclusion from the Revised Amsterdam

[1] Four of these studies comprise unpublished data collected in 2010–2014 and two studies have been published (Stevenson et al. 2013a, 2013b). The published studies did not investigate the moderating roles of working memory or initial strategy-use on the effects of different types of feedback – the research question this paper focuses on.

Children's Intelligence Test (RAKIT, Bleichrodt et al. 1987) or the Standard Progressive Matrices (Raven et al. 1998)). The three intervention conditions presented in this study are: (1) interactive tutoring feedback, (2) multiple-try feedback, or (3) no feedback. Four analogy testing and intervention sessions took place weekly and lasted 15–20 min each. The pretest was administered during the first session, the training sessions took place in the following two sessions, and in the last session the posttest was administered. Prior to the analogy testing sessions the children were also administered an age-appropriate verbal memory test (AWMA listening recall, Alloway 2007; WISC-IV digit span, Wechsler 2003; RAKIT memory span, Bleichrodt et al. 1987). All participants were tested individually in a quiet room at the child's school by educational psychology students trained in the procedure. Each of these studies was approved of by the Leiden University Psychology Research Ethics Committee.

Analogical Reasoning Assessment

AnimaLogica was used to test and train children in analogical reasoning (Stevenson 2012). The figural analogies (A:B::C:?) comprise of 2 × 2 matrices with familiar animals as objects (see Fig. 1). The animals changed horizontally or vertically by color, orientation, size, position, quantity or animal type. The number of transformations – or object changes – provide an indication of item difficulty (Stevenson et al. 2013b). The children were asked to construct the solution to the analogy using drag &

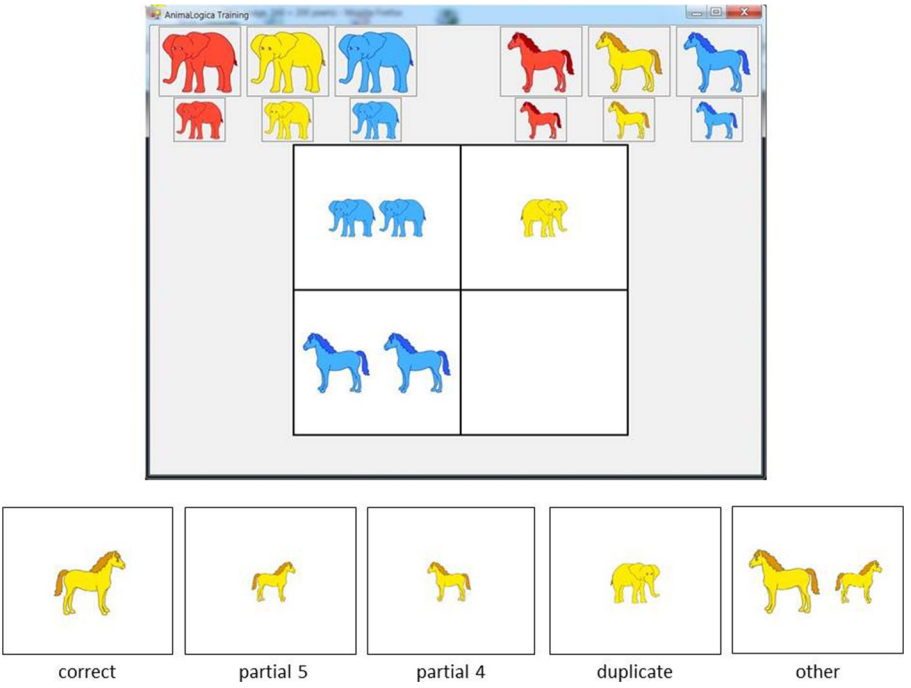

Fig. 1 Example item from AnimaLogica showing the five solution categories. From *left* to *right* the strategies are correct (i.e., all transformations were solved correctly), *partial 5* (where only one of the six transformations was solved incorrectly, in this case *size*), *partial 4* (where two of the six transformations was incorrect, in this case *size* and *orientation*), *duplicate* (a copy of one of the other elements in the analogy) and *other*

drop functions to place animal figures into the empty box in the lower left or right quadrant of the matrix. A maximum of two animals were present in each analogy. These were available in three colors (red, yellow, blue) and two sizes (large, small). The orientation (facing left or right) could be changed by clicking the animal figure. Quantity was specified by the number of animal figures placed in the empty box. Position was specified by location of the figure placed in the box.

The pretest and posttest items were isomorphs in which the items only differ in color and type of animal, but utilize the exact same transformations to ensure the same difficulty level (Stevenson et al. 2013b). The number of items differed per age group (15 items for kindergartners to 24 items for 3rd and 4th graders) and were selected from a pool of 36 items that were expected, based on pilot studies and item difficulty analyses, to cover the entire range of abilities in that age-group (−3 to +3 standard deviations from the mean). At least twelve of the items overlapped with those of the other age groups so that the different tests could be linked using item response theory equating procedures and thereby provide reliable pretest and posttest ability estimates (Embretson and Reise 2000). The internal consistency of each of the test versions was considered very good with $\alpha \geq .90$.

Before each testing or training session two example items were provided with simple instructions on how to solve the analogies. During the training phase the children solved 10 items and received feedback (except if in the control condition); the feedback procedures are described in the next section. During the pretest and posttest items were administered without feedback.

The children's pretest and posttest solutions were categorized into five strategies based on the literature (e.g., Cheshire et al. 2005; Siegler and Svetina 2002) for analyzing strategy-use: (1) correct analogical solutions as correct answer construction, (2) partial analogical with five (of six) transformations solved correctly, (3) partial analogical with four (of six) transformations solved correctly, (4) duplicate non-analogical solutions were copies of the B or C term, and (5) other non-analogical solutions (see Fig. 1). A duplication error was always scored as category 4 – even if the duplicate contained four or five correctly solved transformations.

Feedback Interventions

The *interactive tutoring feedback condition* received training according to the graduated prompts method (Campione and Brown 1987; Resing and Elliott 2011), which consisted of stepwise instructions beginning with general, metacognitive prompts, such as focusing attention, followed by cognitive hints, emphasizing the transformations and solution procedure, and ending with step-by-step scaffolds to solve the problem (see Table 1). The prompts were mostly auditory in nature and accompanied by visual effects such as highlighting relevant aspects of the problem or images (e.g. a brain when the tutor said 'Think.') to support the explanations (see Figs. 2 and 3). A maximum of five prompts were administered. The final prompt, scaffolding, always led to the learner constructing the correct solution. Once the child answered an item correctly the child was asked to explain his/her answer; no further prompts were provided and the next item was administered.

The *multiple-try feedback condition* received auditory and visual feedback on whether or not the outcome was correct and this was repeated until the item was solved

Table 1 Overview of the feedback instructions for the simple repeated feedback condition and the stepwise elaborate feedback condition

	Outcome feedback	Graduated prompts
0	Here's a puzzle with animal pictures. The animals from this box are missing. Can you figure out which ones belong in the empty box?	
1	That's a nice picture but it's not the correct solution. Try again!	Do you remember what to do? Look carefully. Think hard. Now try to solve the puzzle.
2	Oeps, that's not the answer. Give it another try.	This animal picture changes to this one. Here it should change the same way.
3	Hmm, that's not it. Try again, perhaps you can figure it out.	So what changes here (A:B)? Ok remember this one changes the same way.
4	You are clearly doing your best! Try one last time.	See, this picture (A) changes to this one (B) because…
5	This animal belongs in the empty box.	Which animal do you need? Which color? …Size, Quantity, Orientation, Position?

correctly or five attempts were made to solve the item. After the fifth incorrect attempt the correct solution was shown before proceeding to the next item. If a correct solution was found before five attempts then the next item was administered.

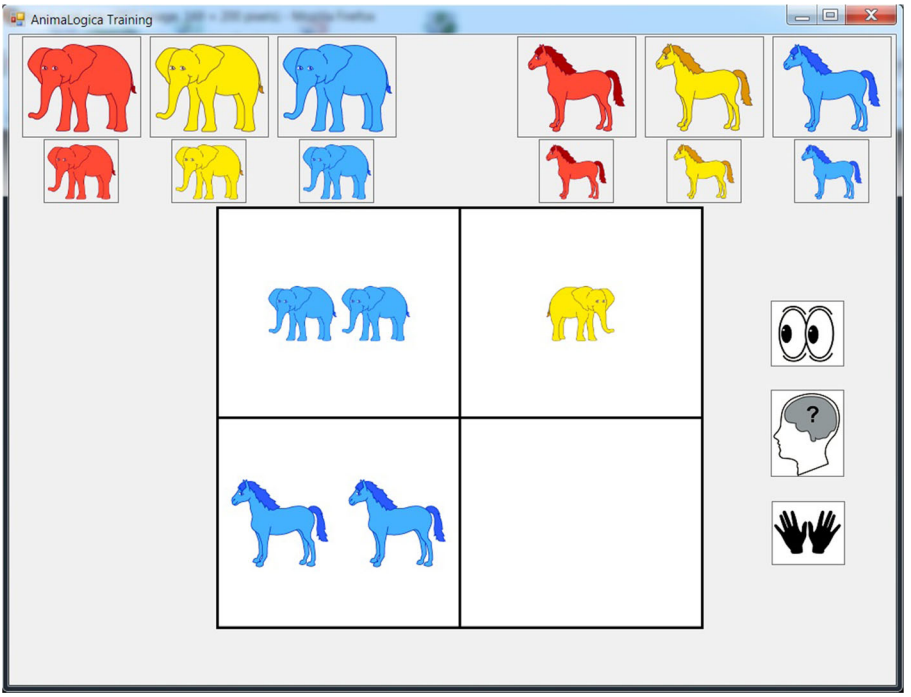

Fig. 2 Depiction of visual effects that emphasize cues from the 2nd prompt in the stepwise elaborate feedback condition to "Look carefully", "Think hard" and then "Try to solve the puzzle" (these are not all shown at once)

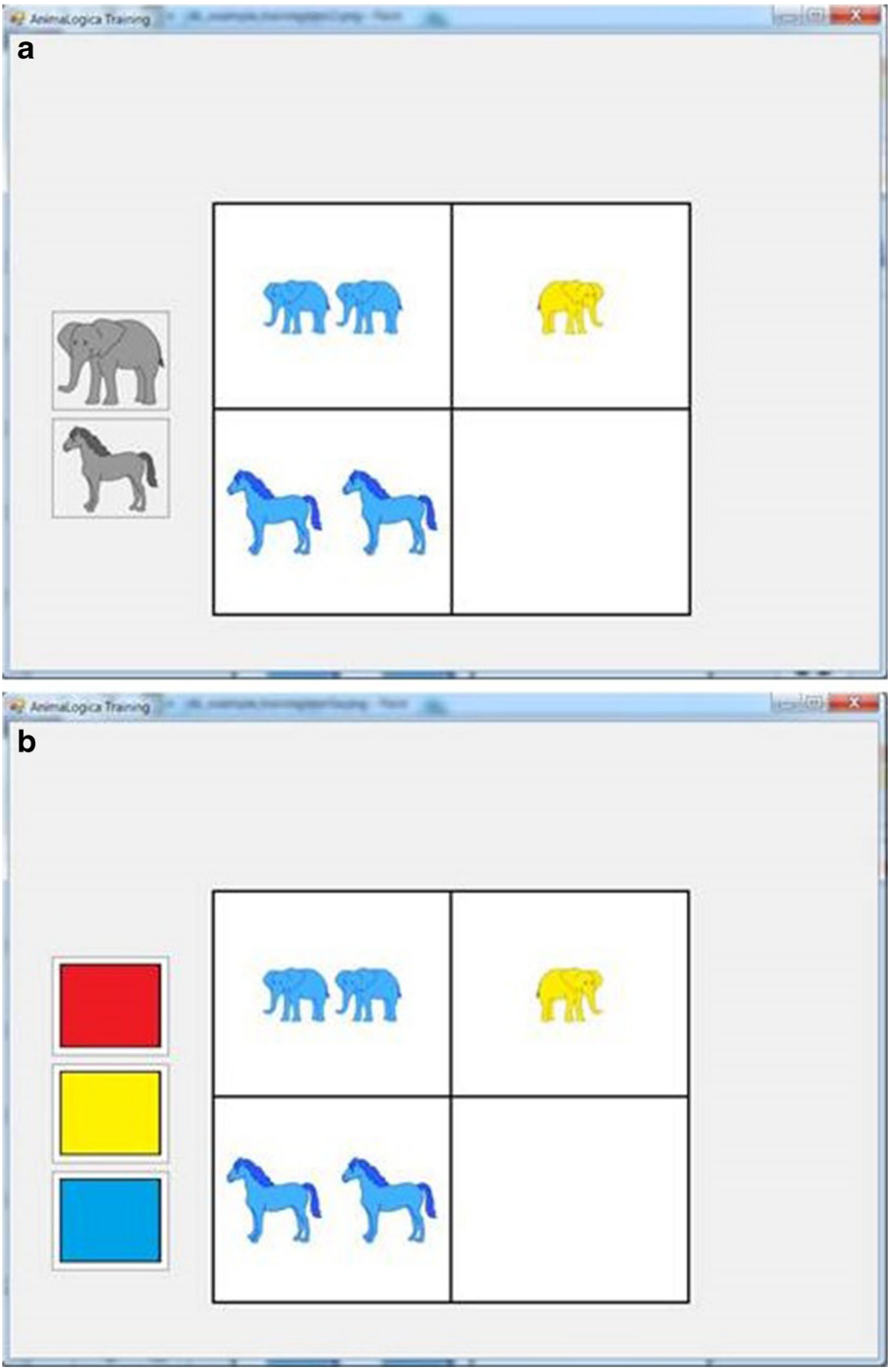

Fig. 3 Visual effects emphasizing prompt 5 where scaffolds are used to solve the puzzle. **a**: "Which animal belongs in the empty box?". **b**: "What color should it be?"

The children in the *control condition* the children received the exact same items during the training sessions as in the other two conditions but did not receive help or feedback in solving them. Therefore, the children only practiced solving the items but were not trained in analogical reasoning.

Statistical Models

Categorizing Initial Ability with Latent Class Analysis

We expected that the children could be classified as different types or subgroups of analogical reasoners, as reflected by the strategies utilized on the pretest items. For instance, one subgroup of children may use predominantly non-analogical strategies such as duplication, while another subgroup of children uses mostly analogical strategies (i.e., correct or partially correct solutions). Latent class analysis (LCA: e.g., Goodman 1974), a cluster analysis technique for categorical observed data, was used to analyze whether such qualitative individual differences were present. In LCA, it is assumed that there is a categorical latent variable that represents latent (unobserved) classes or subgroups of children. The basic latent class model is $f(\boldsymbol{y}) = \sum_{k=1}^{K} P(k) \prod_i P(y_i|k)$. Classesrun from$k$ = 1, . . . , K , and $\boldsymbol{y}$ is a vector containing the observed data: the solution strategy (five categories) on all pretest-items i. LCA is a full-information likelihood approach and can adequately deal with missing data caused by the fact that not all items were administered to all children (missing-by-design). The parameters estimated in LCA are the class sizes $P(k)$ and the conditional probabilities $P(y_i|\, k)$. The latter reflect for each latent class k the probabilities of solving item i with each of the strategies, and serve as the basis for interpreting the classes.

LCAs were carried out with the poLCA package (Linzer and Lewis 2011) available for the statistical computing program R. The BIC-value (Bayesian Information Criterion), a model fit statistic that penalizes the fit of a model with the number of parameters estimated, was used to choose how many latent classes were needed for the current data. Lower BIC-values represent a better trade-off between model fit and parsimony.

Examining Learning over Time with Explanatory Item Response Models

Disentangling the complex changes in ability over time on an individual basis requires complex statistical models. For example, using raw gain scores (posttest minus pretest score) to measure change can lead measurement errors due to the unreliability of the gain score, the regression effect of repeated administration and that the scale units for change do not share constant meaning for test takers with different pretest scores (Embretson and Reise 2000). These problems are potentially solved by placing ability scores for pretest and posttest on a joint interval measurement scale using logistic models such as those employed in item response theory (IRT, Embretson and Reise 2000). In the Rasch model, one of the most simple IRT models, the chance that an item is solved correctly depends on the difference between the latent ability of the learner and the difficulty of the presented item or problem. The Rasch-based gain score provides a good basis for the latent scaling of learning and change because the gain score has the same meaning in terms of log odds (i.e. the logarithm of probability of correct vs. incorrect) across the entire measurement scale

(Embretson and Reise 2000). Therefore, this study applied IRT models to answer the research questions concerning the role of diverse factors in feedback effects on children's change in analogical reasoning.

Each of the hypotheses about the children's performance and change was investigated using model comparison. First a reference model was created and then predictors were added successively to so that the fit of the new model could be compared to the previous (nested) model using a likelihood ratio (LR) test, which assesses change in goodness of fit. The models were estimated using the lme4 package for R (Bates et al. 2014) as described by (De Boeck et al. 2011).

Reference Model The initial reference model was a simple IRT model with random intercepts for both persons and items (pretest and posttest) where the probability of a correct response of person p on item i is expressed as shown in Eq. 1.

$$P\left(y_{pi}=1|\theta_p,\beta_i\right)=\frac{\exp\left(\theta_p-\beta_i\right)}{1+\exp\left(\theta_p-\beta_i\right)} \text{ where } \theta p\sim N(0,\sigma\theta 2) \text{ and } \beta i\sim N(0,\sigma\beta 2) \quad (1)$$

Modeling Learning and Change This study employs repeated testing. In order to account for this effect a session parameter was added to the reference model to represent average change from pretest to posttest. However, this model assumes the effect of retesting to be equal for all children. In order to allow for individual differences in improvement from pretest to posttest a random parameter that allows for the session effect to vary over persons was added. In this model, Embretson's Multidimensional Rasch Model for Learning and Change, the chance that an item is solved correctly (P_{pi}) also depends on the difference between the examinee's latent ability (θ_p) and the item difficulty (β_i) (Embretson and Reise 2000). Yet, the ability is built up in a summation term for the testing occasions m to M, which indicates which abilities (θ_{pm}) must be included for person p on occasion m.

$$P\left(y_{pmi}=1|\theta_{pm},\beta_i\right)=\frac{\exp\left(\sum_m^M\theta_{pm}-\beta_i\right)}{1+\exp\left(\sum_m^M\theta_{pm}-\beta_i\right)} \text{ where } \theta pm\sim N(0,\sigma\theta 2) \text{ and } \beta i\sim N(0,\sigma\beta 2) \quad (2)$$

The initial ability factor, θ_{p1}, refers to the first measurement occasion (i.e. pretest) and the so-called modifiabilities (θ_{pm} with $m > 1$) represents the change from one occasion to the next. In the present model examining pretest to posttest change $M = 2$ and the modifiability θ_{p2} refers to performance change from pretest to posttest.

Modeling Sources of Individual Differences in Learning and Change The formula in Eq. 2 can be extended by including other item or person predictor variables and evaluating their effects on the latent scale (De Boeck and Wilson 2004). Person predictors now include not only initial ability and modifiability, but also factors such as age and working memory scores (Z-scores with M = 0, SD = 1) and are denoted as $_{pmj}$ $(j = 1,\ldots,J)$ and have regression parameters ζ_j. These predictors were successively entered into the null model (see Eq. 1) as follows, with indices i for items, p for persons, j for the person covariate used as a predictor variable.

$$P\left(y_{\text{pmji}} = 1|\theta_{pm1}\ldots\theta_{pmJ},\beta_i\right) = \frac{\exp\left(\sum_m^M\sum_j^J\zeta_j\theta_{pmj}-\beta_i\right)}{1+\exp\left(\sum_m^M\sum_j^J\zeta_j\theta_{pmj}-\beta_i\right)} \text{ where } \theta\text{pmj}\sim N(0,\sigma\theta 2) \text{ and } \beta\text{i}\sim N(0,\sigma\beta 2) \quad (3)$$

Results

Descriptive Statistics

Descriptive statistics of the three feedback conditions are shown in Table 2. The children in the three training conditions differed in average age ($F(2, 996) = 37.91$, $p < .001$, partial $\eta^2 = .07$), but not in average working memory capacity ($F(2, 735) = 2.26$, $p = .11$, partial $\eta^2 = .01$). Age and working memory were not correlated ($r = .004$, $p = .91$).

LCA

We estimated models with 2 and 3 latent classes; the BIC-values were 10,310.9 and 10,311.2 for the 2-class and 3-class models respectively. Based on these BICs we chose to interpret the model with two latent classes or subgroups of children.

Figure 4 shows the conditional probabilities, i.e., for each class the probabilities of using each of the five solution strategies (separate lines) on each of the 24 pretest-items (horizontal axis; ordered with respect to difficulty level/number of transformations). The first class contains 50.7% of the children. Their strategy choice profile is characterized by a large tendency to use a correct AR-strategy on the easier items up to four transformations, and to use either the correct AR-strategy or one of the two partially correct AR-strategies on the more difficult items. Apparently, these children grasp the

Table 2 Descriptive statistics by feedback condition

	Feedback condition			
	Tutoring	Multiple Try	Control	Total
	$N = 427$	$N = 201$	$N = 371$	$N = 999$
	M (SD)	M (SD)	M (SD)	M (SD)
Pretest[a]	.24 (.22)	.31(.26)	.24(.21)	.25 (.23)
Posttest[a]	.48 (.25)	.47(.27)	.31(.25)	.41 (.27)
Training total attempts	18.2 (14.1), range 0–50	16.8 (13.1), range 0–45		
Age[b]	7.17 (1.40)	8.02 (1.52)	6.90 (1.56)	7.25 (1.54)
Memory[c]	-0.04 (1.00)	-0.13 (0.98)	0.06 (0.99)	0.00 (0.99)
Analogical / Non-analogical reasoners (N)	211 / 216	107 / 94	190 / 181	508 /491

[a] proportion correct

[b] in years

[c] test-specific standard scores were converted to z-scores (M = 0, SD = 1); data available for 738 participants

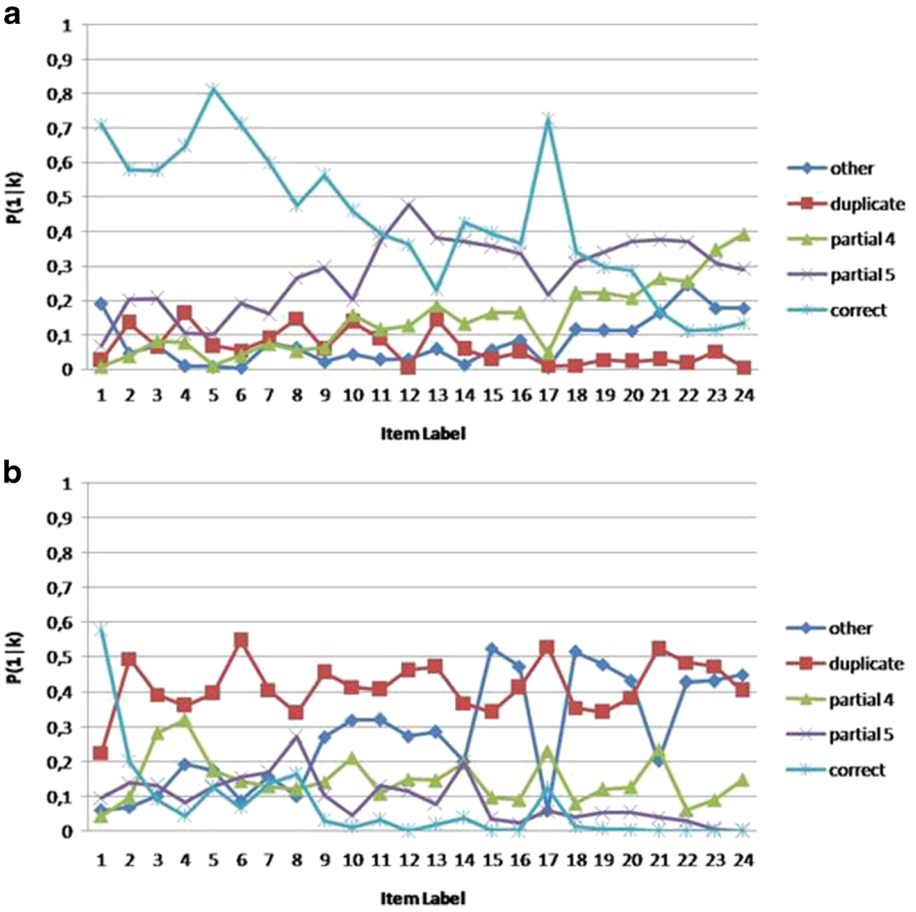

Fig. 4 Depiction of strategy distribution within the two latent class profiles based on pretest strategy-use: analogical reasoners (*top*) and non-analogical reasoners (*bottom*). Analogical reasoners typically apply correct strategies to easier items and partial strategies to more difficult items. Non-analogical reasoners generally apply duplication or other strategies to all but the easiest items

principles of analogical reasoning, but when the items involve more transformations they experience difficulty incorporating them all in their solution. Nevertheless, they appear to have substantial knowledge of analogical reasoning and are therefore labeled 'analogical reasoners'.

The second class (49.3%) consists of children who tend to use duplication on easier items and duplication or 'other' strategies on the more difficult ones. This group has a very low probability of applying correct or partially correct strategies except for the first item. These children appear to have rather limited knowledge of analogical reasoning and were labeled 'non-analogical reasoners'.

Explanatory IRT

The models described in Examining Learning over Time with Explanatory Item Response Models Section were estimated successively; Table 3 displays the outcomes

Table 3 Overview of the estimated multilevel item response theory models

Model	Nested Model	Effects							LR test[b]	
		Fixed[a]	Random over Persons	Random over Schools	Random over Items	AIC	BIC	-LL	*df*	Λ
M0a			Intercept	Intercept	Intercept	37,119	37,153	18,555		
M0b	M0a	+ Item difficulty	"	"	"	37,088	37,130	18,539	1	33.21***
M0c	M0b	+ Age	"	"	"	37,001	37,057	18,497	1	83.62***
M0d	M0c	+ Session	"	"	"	35,172	35,232	17,579	1	1836.16***
M0e	M0d		+ Session	"	"	34,290	34,367	17,136	2	885.73***
M1	M0e	+ Session * Condition	"	"	"	34,160	34,272	17,067	4	137.55***
M2	M1	+ Session * Condition * WM	"	"	"	26,235	26,392	13,099	6	7937.00***
M3	M1	+ Session * Condition * Strategy-Use + WM	"	"	"	25,799	25,965	12,880	7	8374.00***

***$p < .001$

a When adding a fixed effect the * indicates the addition of all main and interaction effects not already present in the model

b The LR-test comprises a comparison between the new model and the previous nested model

of each IRT model building step. The left-most column is the name of the new model. The nested model, i.e. the model to which covariates have been added, is listed in the next column. The next three columns describe which fixed or random effects were added and tested. The model fit indices - AIC and BIC - are displayed in the next columns; for each of the fit indices smaller numbers indicate a better fitting model. The right-most columns contain the values and results of likelihood ratio tests; if these are significant then adding the covariate (s) improved model fit.

Reference Model

M0a represents the reference model and fixed or random effects were added in each successive modeling step. At this step (M0b) we added an item difficulty parameter, i.e. the number of transformations in the analogy, to allow us to control for differential effects of item difficulty (items were tailored to age); furthermore, age was added as a covariate to control for age differences between conditions (M0c). As can be seen in Table 3 both covariates improved model fit.

Modeling Learning and Change

Given the significant improvement in model fit from M0c to M1 we could statistically infer that there was a main effect for the feedback sessions. Furthermore, the inclusion of individual regression lines for performance change from the pretest to posttest was deemed warranted given the improved model fit from M1 to M2, i.e. children differed in how much they improved in analogical reasoning from pretest to posttest.

Research Question 1: Tutoring > Multiple Try > no Feedback

Hypothesis 1 was tested by adding the main and interaction effects of feedback condition to the previous model. The significant model comparison result from M0e to M1 showed us that the different types of feedback had different "change" slopes, i.e. some types of feedback were better at improving analogical reasoning in the children than others. As we can see in Figs. 5 and 6, and in the values reported in Table 4, hypothesis 1 was confirmed: interactive tutoring feedback (ITF) led to greater progression in analogy solving than multiple try feedback (MTF) and multiple try feedback was more effective than no feedback (NF).

A more specific description of the M1 now follows; there were random intercepts for persons ($SD_{ability}$ = 1.38, $SD_{modifiability}$ = 1.23, $r = -.42$), items (SD = .85) and schools (SD = .53). Table 4 reports the estimates of the fixed effects. The effect of the item difficulty (number of transformations in analogy item) was significant ($z = -7.31$, $p < .001$) and indicates that the more transformations an item contains the less likely the participants could solve that item; an increase of 1 transformation lead to .54 decrease in the odds of solving an item correctly. The effect of age was also significant ($z = 9.31$, $p < .001$) and indicates that older children were more likely to solve an item correctly; more specifically, an increase of 1 standard deviation in age leads to 2.31 times increase in the odds of solving an item correctly.

Children in the three conditions did not differ in pretest performance (ITF vs MTF: $z = 0.26$, $p = .79$; MTF vs NF: $z = 0.20$, $p = .84$). Children generally performed better

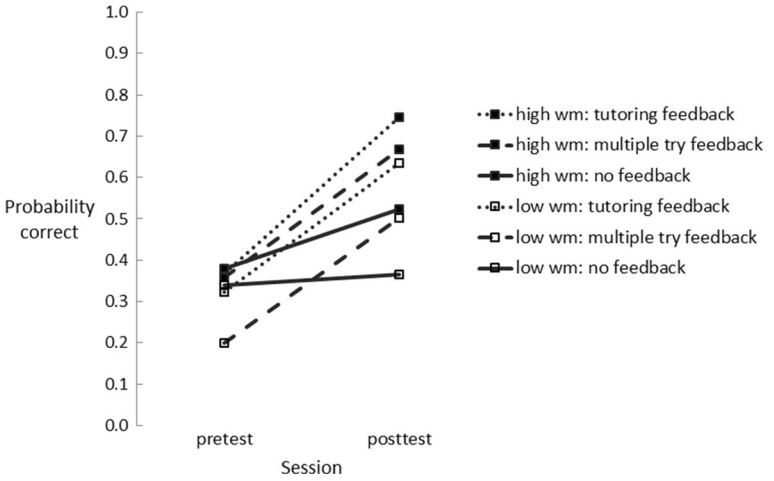

Fig. 5 Model M2 fixed effects depicting the moderating role of working memory (WM) on the effect feedback on children's improvement in analogical reasoning. The probability of solving items correct are shown for children with low (−1.5 SD) and high (+1.5 SD) WM for each of the feedback conditions: tutoring feedback, multiple try feedback or no feedback. Children with higher WM scores had greater learning gains in the tutoring feedback and no feedback conditions, whereas WM did not interact with learning gains in the multiple try feedback condition

on the posttest than the pretest ($z = 11.94$, $p < .001$), with 3.75 greater odds of solving analogies correctly on the posttest. Finally, and most importantly, ITF was more effective than MTF ($z = 3.39$, $p < .001$), where children trained with ITF had 1.57 greater odds of solving a posttest item correctly than children in the MTF group. NF was less effective than MTF ($z = -5.41$, $p < .001$) where children in the MTF group had 2.11 greater odds of solving a posttest item correctly than children in the NF group.

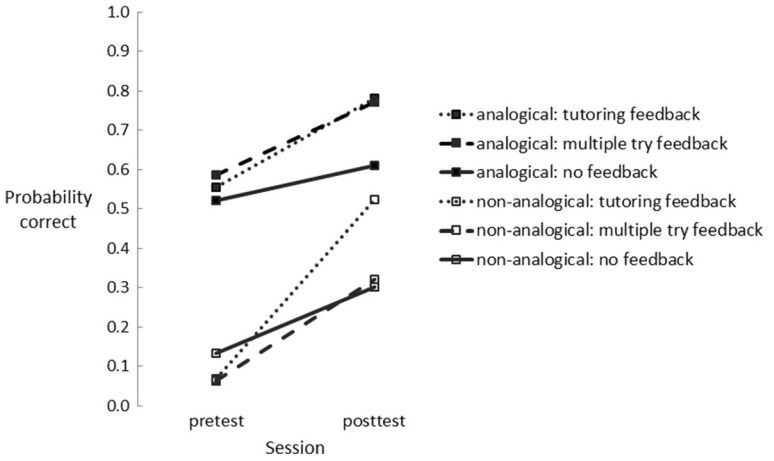

Fig. 6 Model M3 fixed effects showing the effect of strategy-group (analogical versus non-analogical reasoners). Here we see that analogical reasoners initially perform better on the items; however, non-analogical reasoners improve more from pretest to posttest. In both strategy-groups tutoring feedback leads to greater gains than multiple try feedback or no feedback

Table 4 Estimates of random and fixed effects for Models 1–3, testing hypotheses 1–3

	Model 1: Tutoring > Multiple try > No Feedback?				Model 2: Role of working memory?				Model 3: Role of initial strategy-use?			
	B	SE	Z	Odds ratio	B	SE	Z	Odds ratio	B	SE	Z	Odds ratio
Intercept	0.59	0.46	1.29	1.81	0.82	0.43	1.92+	2.27	-1.15	0.43	-2.68**	0.32
Item difficulty	-0.62	0.08	-7.31***	0.54	-0.59	0.08	-7.60***	0.55	-0.58	0.08	-7.54***	0.56
Age	0.84	0.09	9.31***	2.31	0.95	0.09	11.04***	2.60	0.52	0.08	6.89***	1.68
NF Condition (ref = MTF)	0.03	0.16	0.20	1.03	0.10	0.17	0.59	1.10	0.85	0.24	3.49***	2.35
ITF Condition (ref = MTF)	0.04	0.14	0.26	1.04	0.04	0.14	0.30	1.04	0.12	0.19	0.63	1.12
Session (ref = pretest)	1.32	0.11	11.94***	3.75	1.29	0.11	11.23***	3.63	1.97	0.18	11.10***	7.16
Session * NF Condition	-0.75	0.14	-5.41***	0.47	-0.70	0.16	-4.49***	0.49	-0.94	0.26	-3.66***	0.39
Session * ITF Condition	0.45	0.13	3.39***	1.57	0.33	0.15	2.24*	1.39	0.73	0.23	3.13**	2.08
WM					0.54	0.11	4.87***	1.71	0.22	0.04	5.99***	1.25
Session * WM					-0.07	0.12	-0.64	0.93				
WM * NF Condition					-0.42	0.15	-2.81**	0.66				
WM * ITF Condition					-0.41	0.14	-2.90**	0.66				
Session * WM * NF Condition					0.39	0.16	2.45*	1.47				
Session * WM * ITF Condition					0.29	0.15	1.95+	1.34				
Strategy-group (ref = NAR)									3.07	0.19	16.20***	21.55
Session * Strategy-group									-1.10	0.22	-4.95***	0.33
Strategy-group * NF Condition									-1.12	0.27	-4.14***	0.33
Strategy-group * ITF Condition									-0.25	0.23	-1.10	0.78
Session * Strategy-group * NF Condition									0.44	0.31	1.40	1.55
Session * Strategy-group * ITF Condition									-0.54	0.29	-1.88+	0.58

Age and WM scores were converted to z-scores with M = 0 and SD = 1

NF no feedback, MTF multiple try feedback, *ITF* tutoring feedback, *NAR* non-analogical reasoners, *WM* working memory

$^{+}p < .10$, $*p < .05$, $** p < .01$, $*** p < .001$

Research Question 2: Moderating Role of Working Memory

The effect of working memory on children's progression after receiving feedback was examined with model M2. Given the significant improvement in model fit from M1 to M2 we could statistically infer that children with lower working memory scores benefitted differently from the different forms of feedback than children with higher working memory scores. Hypothesis 2, that working memory scores would moderated the interaction between analogy progression and feedback condition, was confirmed. The result of M2 and the direction of this moderation effect is now explained in more detail.

In M2 random intercepts were present for persons ($SD_{ability}$ = 1.32, $SD_{modifiability}$ = 1.31, r = −.54), items (SD = .78) and schools (SD = .45). Table 4 reports the estimates of the fixed effects. Working memory (WM) capacity was related to initial ability (z = 4.87, p < .001), where an increase of 1 standard deviation in the working memory score lead to a 1.71 increase in the odds of solving an item correctly at pretest.

WM scores interacted with feedback condition, indicating that the mean pretest scores of children with the same level of WM differed across feedback conditions; this was lower in the ITF and NF conditions compared to those of the MTF condition (NF: z = −2.81, p < .01; ITF: z = −2.90, p < .01).

As expected, there was a significant three-way interaction between working memory scores, feedback condition and pretest-to-posttest progression. Higher working memory scores were (marginally) related to larger performance gains from pretest to posttest for the ITF and the NF conditions (ITF condition: z = 1.95, p = .051, 1.34 odds increase per +1SD; NF condition: z = 2.45, p < .05, 1.47 odds increase per +1SD). The effect of feedback did not differ for different levels of working memory in the MTF condition: z = −.64, p = .52.

The results are visualized in Fig. 5 which depicts pretest to posttest progression for each condition by two levels of WM scores (WM fixed at +1.5 SD and −1.5 SD). The findings indicate that the greater a learner's working memory the better chance of success s/he has of correctly solving posttest items after training with NF of ITF. WM task performance did not interact with the effect of MTF.

Research Question 3: Moderating Role of Initial Strategy-Use

The difference in performance change from pretest to posttest between the two strategy-groups was added to M1 to form M3. With M3 we examined whether non-analogical reasoners would benefit more from elaborate feedback compared to analogical reasoners after controlling for age and working memory score (hypothesis 3).

In M3 random intercepts were present for persons ($SD_{ability}$ = 0.89, $SD_{modifiability}$ = 1.18, r = −.43), items (SD = .77) and schools (SD = .38). Table 4 reports the estimates of the fixed effects and these are depicted in Fig. 6. There was a difference in pretest performance for non-analogical reasoners between feedback conditions; the children in the NF condition had lower scores than children in the MTF condition on the pretest (z = −4.14, p < .001; .33 lower odds). These initial differences were accounted for in the model and no further pretest performance differences between conditions were apparent.

Analogical reasoners were 21.55 times more likely to solve pretest items correctly than non-analogical reasoners for each of the feedback conditions ($z = 16.20$, $p < .001$). The impact of feedback condition on posttest performance depended on the level of analogical reasoning strategy-use. Non-analogical reasoners had greater learning gains, i.e. improvement from pretest to posttest, than analogical reasoners in the MTF condition ($z = -4.95$, $p < .001$, .33 lower odds). More specifically, if we compare a non-analogical reasoner to an analogical reasoner with the same age and WM score then the non-analogical reasoner has approximately 3 times greater odds of solving an item correctly after training with MTF. The difference in magnitude of the feedback effect for non-analogical versus analogical reasoners for the NF condition did not differ significantly from that of the MTF condition ($z = 1.40$, $p = .16$); however, for the ITF condition there was a marginally steeper slope for non-analogical versus analogical reasoners compared to the MTF condition ($z = -1.88$, $p = .06$). This is shown in Fig. 7 where the non-analogical reasoners improve more from pretest to posttest than analogical reasoners.[2] We can also see that ITF was the most effective feedback form for both non-analogical and analogical reasoners, followed by the MTF condition and then the NF condition; these findings corroborate hypothesis 3.

Discussion

This aim of this paper was to investigate whether children's initial strategies or working memory capacity influenced their learning gains in analogical reasoning after receiving interactive tutoring feedback, multiple try feedback or practice without feedback. The discussion is ordered along the lines of three main findings: (1) interactive tutoring feedback was the most effective intervention; (2) working memory scores moderated feedback effects, where children with higher working memory scores derived the most benefit from interactive tutoring feedback and practice, but benefit from multiple try feedback was unrelated to working memory; (3) feedback effects were influenced by ability level, where children who initially applied non-analogical reasoning strategies improved more than children who used analogical reasoning strategies – especially in the interactive tutoring feedback condition.

Interactive Tutoring Feedback Is more Effective than Multiple Try Feedback

In line with previous findings with adults (e.g., Van der Kleij et al. 2015) and with children (e.g., Cheshire et al. 2005) both simple and elaborate feedback led to greater learning gains than practice without feedback. Furthermore, this study provides evidence that elaborate feedback using a tutoring strategy is more beneficial than simple multiple try feedback for primary school children. Recent reviews and meta-analyses have clearly shown that adults learn more after receiving elaborate feedback than simple feedback such as knowledge of (correct) response (e.g., Van der Kleij et al. 2015). However, children appear to process negative versus positive feedback differently from adults (Eppinger et al. 2009); for example, adults and 11–13 year-olds

[2] By using the IRT logit-based scale with a large range in item difficulty, we were able to rule out that this was caused by ceiling effects.

learned more quickly from negative than positive feedback, whereas 8–9 year olds learned most from positive feedback (Van Duijvenvoorde et al. 2008). Multiple try and interactive tutoring feedback incorporate both negative and positive feedback components, the combination of which has yielded mixed results in children (Barringer and Gholson 1979). It is possible that providing a different kind of simple feedback - e.g., only showing the correct response - would be even more effective for children and this should be examined in future studies.

Possible reasons that elaborate feedback, such as tutoring, is most helpful, also in children, may be that it increases motivation (e.g., Narciss and Huth 2006), leads to better goal-direction, and lowers the cognitive load of the task (Shute 2008). However, each of these possible advantages needs to be tested more thoroughly with young children. Children differ from adults in reward processing; children's intrinsic motivation does not benefit from verbal rewards (e.g., praise), whereas that of adults increases (Deci et al. 1999) - therefore the mechanism behind the effectiveness of elaborate feedback may differ for children. Furthermore, learner characteristics such as working memory and ability level should be taken into account with regard to cognitive load as discussed in the following.

Working Memory Moderates Feedback Effects

Recent research indicates that analogical reasoning skills are closely linked to working memory and executive functions (Richland and Burchinal 2013; Thibaut et al. 2010; Thibaut and French 2016). Therefore we expected a positive relationship between working memory scores and analogy performance. However, the role of working memory in children's analogical reasoning progression after receiving varied forms of training has produced contradictory results (Stevenson et al. 2013a; Stevenson et al. 2013b). Aside from differences in type of feedback there were also differences in age groups between earlier studies; furthermore, statistical power may have been a source of the conflicting results. In this study we were able to investigate this more thoroughly using a large sample size with a broad age range.

As with Fyfe et al. (2015), our findings indicate an interesting interaction between working memory capacity and the effect of feedback on performance gains. The higher a child's working memory score the greater gains made in analogy solving after receiving interactive tutoring feedback. This was also the case with the practice only group (i.e., no feedback condition). However, working memory was not related to gain after training with multiple try feedback. From the perspective of the 'Matthew effect' we might expect that the most capable students gain most from each form of training (Walberg and Tsai 1983) – this is indeed a common finding in research investigating children's benefit from feedback (e.g., Wardlow and Heyman 2016). On the other hand, the right feedback could perhaps reduce cognitive load and the effect of (limited) working memory on performance (e.g., Adam and Vogel 2016; Resing et al. 2017). This was not the case for the interactive tutoring feedback provided in this study and could perhaps be best explained by feedback complexity (e.g., Kulhavy et al. 1985), which included a self-explanation step at the end of the prompting sequence. Self-explanation on its own leads to improved performance and likely deeper learning in the context of analogical reasoning (Cheshire et al. 2005; Siegler and Svetina 2002). However, interactive tutoring feedback without self-explanation has led to greater learning gains in the math

domain than multiple try feedback (Narciss and Huth 2006). In our case, learning gains from multiple try feedback was not linked to working memory capacity, therefore our results are puzzling because it leads to the conclusion that tutoring feedback plus self-explanation does not provide enough additional support beyond that of multiple try feedback for children with lower working memory scores. Perhaps immediately eliciting self-explanations overloads the system – at least for children with below average working memory. Future research could perhaps best disentangle the interactions between task complexity, feedback type (with and without self-explanation) and working memory in an experimental design in which performance and change are examined on a trial-by-trial basis where item difficulty can be manipulated to tax the executive system in varying degrees during feedback interventions.

Ability Level Based on Strategy-Use Profile Moderates Feedback Effects

Initial strategy-use was used as an indicator of analogical reasoning skills and using latent class analysis children were categorized as non-analogical or analogical reasoners. Both groups benefitted relatively more from tutoring feedback than multiple try feedback and the least from practice alone. However, non-analogical reasoners (after accounting for differences in age and working memory) benefitted far more from each type of feedback than analogical reasoners – children who generally solved analogies correctly or only made processing mistakes (e.g., picked the wrong color or orientation, but otherwise solved the task correctly). These findings are largely in line with previous finding with children (Narciss and Huth 2006) and theoretical models of feedback effects based on a long line of research concerning the role of prior knowledge and abilities (e.g., Bangert-Drowns et al. 1991; Mason and Bruning 2001), where lower ability learners benefit most from immediately receiving the correct response and also elaboration in the form of concrete and directive scaffolds or hints (Shute 2008). In this study analogical reasoners, i.e. higher ability students, also benefitted relatively more from elaborate (tutoring) feedback than simple (multiple-try) feedback, whereas earlier studies did not always find a difference (e.g., Clariana 1990; Hanna 1976). However, this may be because the present study focused on children, whose executive functions and analogical reasoning skills are still developing (Thibaut and French 2016).

The strategy-profiles appear to be a good way to identify which children have little understanding of the task at pretest and interactive tutoring feedback in the form of graduated prompts clearly "taught" the children how to solve analogies. As such graduated prompts may be an especially useful elaborate feedback procedure in digital learning environments to help children with low initial ability and little understanding of the task perform successfully.

Methodological Implications

A unique component of this study was that we utilized an item response theory (IRT) model based on Embretson's Multidimensional Rasch Model for Learning and Change and extended this with multilevel (Pastor 2003) and explanatory (c.f. De Boeck and Wilson 2004) components. The results are likely similar to those using a combination of classical test theory sum scores and standard linear regression techniques. However, the application of multilevel IRT provided three advantages that would not have been possible with other

techniques. First, the classically measured gain score (posttest correct minus pretest correct) is unreliable and can lead to spurious findings, but IRT change scores are not (e.g., Embretson and Reise 2000) and using these allowed for reliable estimates of learning from feedback. Second, the dataset we used was a combination of smaller experiments with slightly different item sets due to age differences among the participants. With IRT these datasets could be linked (equated) to ensure that the latent trait measured was placed on the same scale for each of the datasets at both pretest and posttest (Embretson and Reise 2000). Third, it was possible to account for random school and experiment differences by including an additional level in the implemented hierarchical model.

Limitations

A few limitations deserve mention when drawing conclusions about the role of feedback on children's change in analogical reasoning based on these results. First, two possible confounders may have led to the reported advantages of tutoring feedback. The first confounder is that the interactive tutoring feedback also incorporated a self-explanation step. Self-explanation in itself is an effective training tool (Roy and Chi 2005). Cheshire et al. (2005) concluded in their study disentangling the effects of feedback and self-explanation that feedback was essential in improving children's analogical reasoning; however, as mentioned above, the combination of tutoring feedback and self-explanation may be relatively more taxing for children than only tutoring feedback and have diminished how effective this elaborate form of feedback is for children. The second confounder is time-on-task, which was not accounted for in the analyses. The children who received tutoring feedback inevitably spent more time working on the analogies than the children in the multiple try condition and this was also greater than the time spent on task by the children in the no feedback group. Furthermore, time-on-task was related to how many errors a child made in analogy solving as feedback was only given after incorrect trials. It is possible that time-on-task can partly explain the differences in feedback effects. However, these differences were a matter of minutes as the children were trained on only ten items spread over two sessions, so time-on-task is unlikely the main cause of differences in feedback effectiveness.

Second, the training the children received was of low intensity with two training sessions; perhaps the results would differ if children were trained for more sessions. A study with four or more measurement moments would reveal more fine-grained information on how feedback affects children's change in figural analogy solving. Third, the current study does not provide information about transfer (e.g., Stevenson et al. 2013a) or long-term effects of the training. Including transfer tasks, such as inductive reasoning seriation, at posttest and retesting after a three month period would provide additional information on the effects of the different feedback paradigms, which might be especially informative for educational purposes. Fourth, motivation effects were not assessed; however, motivation plays an important role in training effectiveness and should form part of the evaluation of different feedback interventions (e.g., Narciss 2013). Finally, including additional school and child characteristics (e.g., school quality, SES) would improve the generalizability of these results.

Conclusion and Future Directions

The main conclusion is that interactive tutoring feedback, an elaborate form of feedback presently implemented using graduated prompting techniques, appears to be the advisable form of feedback in advancing children's analogical reasoning. However, how effective elaboration is clearly depends on the learner characteristics assessed in this study: working memory and ability based on strategy-use profile. The most important exception was that children with lower working memory scores benefitted similarly from multiple try feedback and interactive tutoring feedback.

Given the great potential of digital learning environments to provide feedback tailored to an individual's instructional needs an important task is to determine which factors affect feedback processing and how these interact with other learner characteristics so that we can optimize feedback provision and thus learning. On the one hand, (meta-analyses of) randomized pretest-training-posttest control experiments that contrast effects of different types of feedback and explore sources of individual differences herein provide essential information on which factors could be used to optimize feedback. However, investigating the effects of specific feedback prompts on a trial-by-trial basis (e.g., Goldin et al. 2012; Narciss et al. 2014a) and how these interact with learner characteristics (e.g., working memory) and task performance (e.g., strategy-use) using item response theory models is a promising next step towards refining guidelines to provide optimal feedback in digital learning environments. This study contributes to the growing literature on how to provide personalized feedback, i.e. feedback that adapts to the learner's instructional-needs.

Acknowledgements The author greatly appreciates the critical comments and helpful suggestions of the anonymous reviewers and this special issue's editors. Thank you also to Marian Hickendorff for help with the latent class analysis and Paul de Boeck, Willem Heiser and Wilma Resing as supervisors of the author's initial data collection and analysis.

References

Adam, K. C., & Vogel, E. K. (2016). Reducing failures of working memory with performance feedback. *Psychonomic Bulletin & Review, 23*(5), 1520–1527.

Alloway, T. P. (2007). Automated working memory assessment. London: Pearson Assessment.

Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology, 87*, 85–106. doi:10.1016/j.jecp.2003.10.002.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Barringer, C., & Gholson, B. (1979). Effects of type and combination of feedback upon conceptual learning by children: implications for research in academic learning. *Review of Educational Research, 49*(3), 459–478.

Bates, D., Maecheler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *Journal of Statistical Software*, *R package.* Retrieved from http://cran.r-project.org/package=lme4.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). *Handleiding bij de Revisie Amsterdamse kinder Intelligentie test [manual of the revised Amsterdam child intelligence test]*. Lisse: Swets & Zeitlinger.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: an interactional approach to evaluating learning potential* (pp. 82–109). New York: Guilford Press.

Cheshire, A., Ball, L. J., & Lewis, C. N. (2005). Self-explanation, feedback and the development of analogical reasoning skills: microgenetic evidence for a metacognitive processing account. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 435–441).

Clariana, R. B. (1990). A comparison of answer-until-correct feedback and knowledge of correct-response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction, 17*(4), 125–129.

Dagher, Z. R. (1995). Review of studies on the effectiveness of instructional analogies in science education. *Science Education, 79*(3), 295–312.

De Boeck, P. A. L., & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach*. New York: Springer.

De Boeck, P. A. L., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item reponse models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–27.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum Publishers.

Eppinger, B., Mock, B., & Kray, J. (2009). Developmental differences in learning and error processing: evidence from ERPs. *Psychophysiology, 46*(5), 1043–1053.

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: the causal role of prior knowledge. *Journal of Educational Psychology, 108*(1), 82.

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2015). When feedback is cognitively-demanding: the importance of working memory capacity. *Instructional Science, 43*(1), 73–91.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Goldin, I. M., & Carlson, R. (2013). Learner Differences and Hint Content. In *Proceedings of the International Artificial Intelligence for Education Conference (AIED)*, pp. 522–531.

Goldin, I. M., Koedinger, K. R., & Aleven, V. (2012). Learner Differences in Hint Processing. In *Proceedings of the 5th International Conference on Educational Data Mining (EDM)* pp. 956–960.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231. doi:10.1093/biomet/61.2.215.

Goswami, U. (1991). Analogical reasoning: what develops? A review review of research and theory. *Child Development, 62*, 1–22.

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. *Journal of Educational Research, 69*(5), 202–205.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254.

Kramarski, B., & Zeichner, O. (2001). Using technology to enhance mathematical reasoning: effects of feedback and self-regulation learning. *Educational Media International, 38*(2–3), 77–82.

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedbackcomplexity and corrective efficiency. *Contemporary Educational Psychology, 10*(3), 285–291.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: an R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29.

Mason, B. J., & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. Center for Instructional Innovation, University of Nebraska–Lincoln. Retrieved February 15, 2007, from http://dwb.unl.edu/Edit/MB/MasonBruning.html.

Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental Psychology, 51*(3), 214–228.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. G. van Merrienboer, & M. P. Driscoll (Eds.) (pp. 125–143). Mahaw, NJ: Lawrence Erlbaum Associates.

Narciss, S. (2013). Designing and evaluating intelligent tutoring feedback strategies for digital learning environments on the basis of the interactive intelligent tutoring feedback model. *Digital Education Review, 23*, 7–26.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 181–195). Münster: Waxmann Verlag GmbH.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related intelligent tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction, 16*(4), 310–322. doi:10.1016/j.learninstruc.2006.07.003.

Narciss, S., Sosnovosky, S., & Andres, E. (2014a). Adapting intelligent tutoring feedback strategies to motivation. *Open Learning and Teaching in Educational Communities*, 8719. doi:10.1007/978-3-319-11200-8.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014b). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education, 71*, 56–76. doi:10.1016/j.compedu.2013.09.011.

National Academy of Education (2013). *Adaptive educational technologies: tools for learning and for learning about learning*. Washington DC: National Academy of Education.

Pastor, D. A. (2003). The use of multi-level item response theory modeling in applied research: an illustration. *Applied Measurement in Education, 16*(3), 223–243. doi:10.1207/S15324818AME1603_4.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales. Section 4: the advanced progressive matrices*. San Antonion: Harcourt Assessment.

Resing, W. C. M., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: measuring children's change in strategy use with a series completion task. *The British Journal of Educational Psychology, 81*(Pt 4), 579–605. doi:10.1348/2044-8279.002006.

Resing, W., Stevenson, C. E., & Bosma, T. (2012). Dynamic testing: measuring inductive reasoning in children with developmental disabilities and mild cognitive impairments. *Journal of Cognitive Education and Psychology, 11*(2), 159–178.

Resing, W. C., Bakker, M., Pronk, C. M., & Elliott, J. G. (2017). Progression paths in children's problem solving: the influence of dynamic testing, initial variability, and working memory. *Journal of Experimental Child Psychology, 153*, 83–109.

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science, 24*(1), 87–92. doi:10.1177/0956797612450883.

Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction, 22*(1), 37–60.

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271–286). New York City: Cambridge University Press.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. doi:10.3102/0034654307313795.

Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: comparing short-term and long-term change. *Child Development, 73*(3), 793–809 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12038552.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112*(1), 80–116. doi:10.1037//0096-3445.112.1.80.

Stevenson, C. E. (2012). *Puzzling with potential: dynamic testing of analogical reasoning in children*. Amsterdam: Leiden University.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. M. (2013a). Working memory as a moderator of training and transfer of analogical reasoning in children. *Contemporary Educational Psychology, 38*(3), 159–169. doi:10.1016/j.cedpsych.2013.02.001.

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & Boeck, P. A. L. D. (2013b). Intelligence explanatory item response modeling of children ' s change on a dynamic test of analogical reasoning. *Intelligence, 41*(3), 157–168. doi:10.1016/j.intell.2013.01.003.

Stevenson, C. E., Heiser, W. J., & Resing, W. C. (2016). Dynamic testing of analogical reasoning in 5-to 6-year-olds multiple-choice versus constructed-response training items. *Journal of Psychoeducational Assessment, 24*(6), 550–565. doi:10.1177/0734282915622912.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning & Instruction, 4*(4), 295–312.

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: a developmental investigation with eye-tracking. *Cognitive Development, 38*, 10–26.

Thibaut, J.-P., French, R., & Vezneva, M. (2010). The development of analogy making in children: cognitive load and executive functions. *Journal of Experimental Child Psychology, 106*(1), 1–19. doi:10.1016/j.jecp.2010.01.001.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes a meta-analysis. *Review of Educational Research, 85*(4), 475–511.

Van Duijvenvoorde, A. C. K., Zanolie, K., Rombouts, S. A. R. B., Raaijmakers, M. E. J., & Crone, E. A. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *The Journal of Neuroscience, 28*(38), 9495–9503.

Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359–373.

Wardlow, L., & Heyman, G. D. (2016). The roles of feedback and working memory in children's reference production. *Journal of Experimental Child Psychology, 150*, 180–193.

Wechsler, D. (2003). *Wechsler intelligence scale for children-fourth edition. Administration and scoring manual*. San Antonio: Harcourt Assessment, Inc..